Danmarks Tekniske Universitet

Skriftlig prøve, den 21. januar 2013 Side 1 af 22 sider

Kursus navn: Bioinformatik 2

Kursus nummer: 27634

Hjælpemidler: ALLE HJÆLPEMIDLER

Varighed: 4 timer

Navn:

Studienummer:

Unix and databases (25%)

1) Log in to the server and copy the following file to your own home directory: /usr/opt/www/pub/CBS/courses/bioinformatics_it_and_health/2012/bioinformatics2/HLA-A0201.values.log50k
    a) How many lines are in the file.

7064

    b) What is the value assigned to the peptide CHATLTHRL

0.0846866

    c) What is the line number of the peptide CHATLTHRL

555

    d) How many of the peptides in the file have an assigned value higher than 0.5

1960

e) How many of the peptides in the file have an L at position 2?

1961

f) Is any of the peptides in the file duplicated? How did you check that?

No:

```
cat HLA-A0201.values.log50k | wc -l
cat HLA-A0201.values.log50k | gawk '{print($1)}'| sort -u | wc -l
```

Machine Learning, Performances and Evaluation (20%)

1) Are scoring using position specific scoring matrices (PSSMs) to consider as a 'linear' method?

Yes

2) What are the benefits of using more complicated methods such as artificial neural networks (ANNs) or Support vector machines (SVMs)

These methods can capture higher order correlations e.g. solve XOR problems.

3) A hypothetical genome based prediction method have been developed to predict if a bacteria is a pathogen or not. In a set of 122 bacteria 33 were predicted to be a pathogen. Of these 27 could be verified to be pathogens in database searches. 5 strains predicted not to be pathogens were also stated as pathogens in the database. What is the Matthews correlation coefficient of the method?

$$Sens = \frac{TP}{AP}$$

$$Spec = \frac{TN}{AN}$$

$$CC = \frac{TP \cdot TN - FN \cdot FP}{\sqrt{PP \cdot AN \cdot AP \cdot PN}}$$

TP = 27
FP = 6
TN = 84 (122 - 33 - 5)
FN = 5

PP = 33
AP = 32 (27 + 5)
PN = 89 (122 -33)
AN = 90 (122 - 32)

CC = 0.7695

4) Get the files http://www.cbs.dtu.dk/courses/bioinformatics_it_and_health/2012/bioinformatics2/trainset.pep and http://www.cbs.dtu.dk/courses/bioinformatics_it_and_health/2012/bioinformatics2/evalset.pep
You must save as a text or source file. Be sure that the saved files is in raw text.
    a) Use easypred to train a PSSM training with trainset.pep and evaluating with eval.pep. Leave all settings at default. What is the Aroc value and the Pearson correlation on the evaluation set.

Evaluation of predictions
Pearson coefficient for N= 100 data: 0.73403
Aroc value: 0.96373

    b) Change the weight on prior (weight on pseudocounts) to 200 and repeat the training/evaluation. Which values do you get now?

Evaluation of predictions
Pearson coefficient for N= 100 data: 0.73806
Aroc value: 0.96640

c) Explain the effects of pseudocounts.

Pseudocounts gives a probability of finding all amino acids at a given position using the observed amino acids and a substitution matrix (eg. BLOSUM62).

This will help giving weights to non observed amino acids that are not observed because of lack of data.

The weight is set so the effect of the pseudocounts will be 'diluted' when enough data are present so that the real space are expected to be covered by actual observations.

d) Now train using artificial neural networks using 5x cross validation and 5 hidden units. What evaluation performance do you get?

Evaluation of predictions
Pearson coefficient for N= 100 data: 0.81910
Aroc value: 0.93867

e) What is the purpose of cross validation?

Cross validation is a method for evaluating the performance of a prediction method splitting the data in training and evaluation sets. In this way all the available data can be used for training and an independent performance still be calculated.

NGS (15%)

You are responsible for testing for bacterial outbreaks in the Copenhagen area. The last few days more and more people are being hospitalized with severe bacterial infections. The suspected cause is a bacteria which have been isolated and must be sequenced. You would like to do the sequencing using an Illumina sequencer.

1) Would you expect to obtain reads all of the same length or with variable lengths and where do you expect your base quality to be poorest in the reads?

Svar: I en Illumina sekventering er reads altid lige lange, da der sættes en base på hvert read af gangen. Kvaliteten af base-calls er altid dårligst i 3' enden.

2) Explain why the most common errors in 454 sequencing will be insertions and deletions (indels). Why do Illumina obtained data not contain the same type of errors?

Svar: I 454 reads kan der blive indsat flere af de samme baser på én gang, hvilket giver et større signal, men signalet er ikke ligefrem proportionalt med antallet af baser. Derfor kan der nogle gange bliver sat én base for meget eller én base for lidt ind. For Illumina data bliver baserne sat ind én af gangen.

3) Your DNA are sequenced as paired end. Explain what this means and what the benefits are in relation to your analysis?

Svar: Paired end betyder at vi har reads fra det samme DNA-fragment, men fra hver ende af det. Når vi kender længden af DNA-fragmentet og derved hvor langt fra hinanden de to reads bør være, kan vi bruge den information til at lave mere præcis read mapping og til eg. *de novo* assembly.

4) The genome turn out to be approximately 4 Mbases. Calculate how many paired end reads you need in order to obtain a 50x coverage of the genome. Your reads are in average 100 nucleotides.

Svar: C = N * (L/G), hvor C = coverage, N = number of reads, L= read length, G = genome size.
Dvs. N = C*G/L = 50*4M /100 = 2M reads

SOAP and web services (15%)

1) Explain the main differences between RESTful web services and SOAP.

2) Which of the immunology related web servers at www.cbs.dtu.dk/services can be called using SOAP?

NetChop (x2)
NetCTL
NetMHCcons (x2)

3 (5 is accepted because of duplicates)

Project related questions (25%)

1) Which positions in a 9mer peptides are in relation to MHC binding designated ‘anchor positions’ and why?

Positions 2 and 9 are most important in bindimg as these bind in deeper pockets in the MHC and thus contribute more to the binding energy

2) If you saved the output from SignalP as a raw text file you would be able to find the proteins with predicted signal peptide using a few unix commands. How would you do?

3) How large a fraction of the total number of predicted binders are predicted to bind to HLA-A*02:01 encoded MHC? Comment the answer.

4) Do yoy find any relation between the length of your investigated proteins and the number of predicted binding peptides in that protein? Comment the answer.

5) Below is a table with the number of genes in different organisms including bacteria. Under the assumption that one gene encodes one protein, how does the number fiot with the number you found for your organism? Comment the answer.

TABLE 17.1

Representative Sequenced Genomes

ORGANISM	HAPLOID GENOME SIZE (Mb)	NUMBER OF GENES	PROTEIN-CODING SEQUENCE
Bacteria			
M. genitalium	0.58	485	88%
H. influenzae	1.8	1,738	89%
E. coli	4.6	4,377	88%
Yeasts			
S. cerevisiae	12.5	5,770	70%
S. pombe	12.5	4,929	60%
Plants			
A. thaliana	115	28,000	25%
Rice	390	37,544	12%
Animals			
C. elegans	100	18,427	25%
D. melanogaster	123	13,379	13%
Pufferfish	342	27,918	10%
Chicken	1,130	25,000	3%
Human	3,300	24,000	1.2%

Mb = millions of base pairs